An Integrated Docking Pipeline for the Prediction of Large-Scale Protein-Protein Interactions

Xin Hu, Michael Lee, Kamal Kumar, and Anders Wallqvist
Biotechnology HPC Software Applications Institute, Telemedicine and Advanced Technology Research Center, US Army Medical Research and Materiel Command (MRMC), Ft. Detrick, MD
{xhu, kamal, awallqvist}@bioanalysis.org; michael.scott.lee@us.army.mil

Abstract

Knowledge of the three-dimensional (3D) structures of protein complexes provides a fundamental understanding of biological systems, as well as novel insights for antimicrobial drug and vaccine design. Protein-protein docking is used to predict the 3D structures of protein complexes from their components in silico. In this study, we developed a protein-protein docking pipeline (PPDP) that integrates a variety of state-of-the-art protein docking and structure prediction techniques, providing a systematic platform to predict large-scale protein-protein interactions (PPIs). The PPDP is deployed on high performance computing (HPC) clusters, thus enabling Department of Defense scientists to harness HPC resources to investigate the PPIs of biowarfare agents for the development of countermeasures. We applied the PPDP to investigate the binding interactions of Yersinia effector proteins with their chaperone and the underlying specificity of chaperone/effector interactions.

1. Introduction

Protein-protein interactions (PPIs) underlie many basic biological processes. Knowledge of the three-dimensional structures of protein-protein complexes is of fundamental importance for the understanding of biological systems, e.g., in host-pathogen interaction networks, and thus is extremely valuable to antimicrobial drug and vaccine design. Rapid advances in gene sequencing technologies and genome-wide protein structure determination have made *in silico* protein-protein docking a promising approach for the systematic determination and structural characterization of PPIs at the atomic level (Ritchie, 2008).

Protein-protein docking algorithms are designed to predict the three-dimensional (3D) structure of a protein-protein binding complex from its component structures *in silico*. A number of protein docking programs have been developed in the past decades. As representatives, ZDOCK and RosettaDock are two widely-used programs. ZDOCK uses fast Fourier transform (FFT) to globally search rigid-body transformations of two proteins, and is amenable to large-scale decoy generation (Chen, 2003). In contrast, RosettaDock uses Monte Carlo-based searching and local perturbations with structural refinement (Gray, 2003). Although it has been successfully applied for the prediction of protein complexes that agree well with experiments, major challenges still remain with protein docking, especially as it relates to flexible proteins and induced conformational changes upon binding. To improve docking accuracy, a variety of post-processing approaches have been developed. These typically include clustering, hotspot filtering, re-scoring, and structural refinement with minimization and/or molecular dynamics simulations (Andrusier, 2008).

An important application of protein-protein docking is to predict large-scale PPIs at the genomic level (Vajda, 2002). In this case, a large number of proteins are cross-docked and the true interacting partners are identified based on the docked scores. This poses a more challenging problem with protein docking as it requires not only the prediction of the correct binding of two proteins but also the ability to distinguish true binding partners from a large pool of non-binders. In practice, this also demands tremendous computing resources and data mining processes. Although several public Web servers of protein docking are available, they are generally restricted to a single protein-protein docking simulation and not suitable for large-scale PPI prediction studies.

In this study, we developed a protein-protein docking pipeline by integrating various state-of-the-art protein docking programs and structural refinement techniques for systematic prediction of large-scale

978-0-7695-4392-5/10 $26.00 © 2010 IEEE
DOI 10.1109/HPCMP-UGC.2010.18

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2010		**00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
An Integrated Docking Pipeline for the Prediction of Large-Scale Protein-Protein Interactions	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Army Medical Research and Materiel Command,Biotechnology High Performance Computing Software Applications Institute,Telemedicine and Advanced Technology Research Center,Fort Detrick,MD,21702**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
2010 DoD High Performance Computing Modernization Program Users Group Conference, 14-17 Jun, Chicago, IL.

14. ABSTRACT
Knowledge of the three-dimensional (3D) structures of protein complexes provides a fundamental understanding of biological systems, as well as novel insights for antimicrobial drug and vaccine design. Protein-protein docking is used to predict the 3D structures of protein complexes from their components in silico. In this study, we developed a protein-protein docking pipeline (PPDP) that integrates a variety of state-of-the-art protein docking and structure prediction techniques, providing a systematic platform to predict large-scale protein-protein interactions (PPIs). The PPDP is deployed on high performance computing (HPC) clusters, thus enabling Department of Defense scientists to harness HPC resources to investigate the PPIs of biowarfare agents for the development of countermeasures. We applied the PPDP to investigate the binding interactions of Yersinia effector proteins with their chaperone and the underlying specificity of chaperone/effector interactions.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**5**	

protein complexes. Two docking modules, based on ZDOCK and RosettaDock, were designed in combination with a number of post-processing approaches to improve the docking accuracy. The pipeline was automated and deployed on high performance computing (HPC) clusters with a Web-based graphical user interface (GUI), providing a useful tool for scientists in the Department of Defense (DoD) to harness HPC resources to study the PPIs of biowarfare agents for the development of countermeasures.

2. Methodology

The Protein-Protein Docking Pipeline (PPDP) integrates freely-downloadable software components from various academic and government research laboratories. The pipeline comprises two basic docking modules: 1) a ZDOCK-based docking module (ZDM) that generates a large number of "low-resolution" decoys of binding complexes and 2) a RosettaDock-based docking module (RDM) for "high-resolution" docking with flexible local perturbations and structural refinement. At the post-processing stage, a number of approaches are integrated with each docking module. These include biochemical information filtering, pair-wise root mean square deviation calculations and clustering using the MMTSB tool (Feig, 2004), energy minimization using the Amber molecular modeling package (Case, 2008), and re-scoring using a variety of objective functions, i.e., ZRANK (Pierce, 2007), DFIRE (Zhang, 2004), EMPIRE (Liang, 2007), and MM-PB/GBSA (Onufriev, 2000).

As an integrated pipeline, the PPDP provides a systematic platform to design and optimize different docking protocols for the applications of PPI studies. An efficient hierarchical docking module is implemented, which generates a large number of decoys with the ZDM, followed by subsequent local perturbation and the structural refinement of top-scoring models with the RDM. In addition, the PPDP can be integrated with other Biotechnology HPC Software Application Institute (BHSAI) pipeline tools, such as the Protein Structure Prediction Pipeline (PSPP) for homology model prediction (Lee, 2009), or the Automated Protein Ensemble Generator (APEG) for ensemble generation. A workflow of the integrated PPDP is shown in Figure 1.

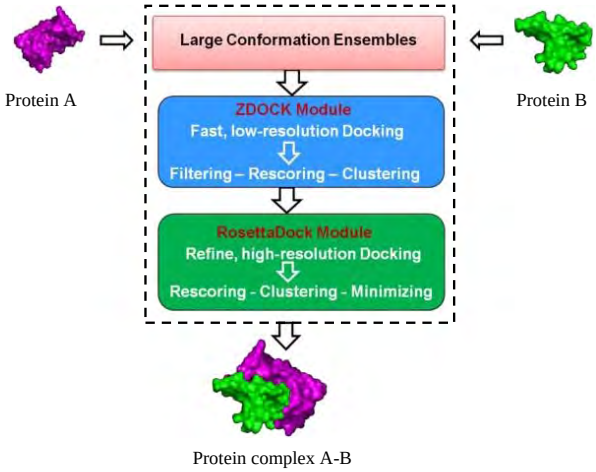

Figure 1. A workflow of the Protein-Protein Docking Pipeline (PPDP)

The PPDP is specifically designed for an HPC system to study large-scale PPI networks. Different from publicly available servers, the PPDP can process a large dataset of protein complexes in parallel. The inputs for the PPDP are two lists of proteins (receptor and ligand), and the output is a score matrix of top-ranked models of each docking partner. The PPDP utilizes the native "Job Array" feature of the queuing system to efficiently manage multiple job submissions and execution. Currently, the PPDP is deployed on the HPC cluster "MANA" at the Maui HPC Center and "MJM" at the US Army Research Laboratory (ARL) DoD Supercomputing Resource Center (DSRC) using the PBS queuing system. In general, it takes ~1 to 2 hours to dock one protein complex with the single-core ZDOCK, whereas the single-core RosettaDock needs more than 5 days to finish one protein-docking task. Therefore, we assessed the performance of the parallelized RosettaDock_MPI on HPC clusters. Figure 2 shows the speed-up curve tested on the MJM cluster. It can be seen that the parallelized RosettaDock achieves near-linear speed-up up to 64 cores.

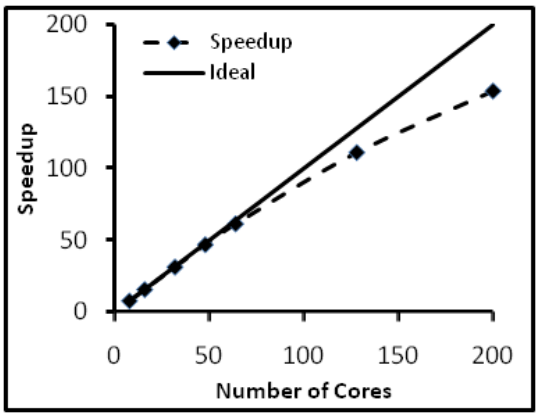

Figure 2. The speed-up curve for the PPDP

To make the PPDP accessible to a broader community of scientists in DoD, we developed a Web-based, user-friendly interface. The GUI uses the User Interface Toolkit to allow authorized personnel to access HPC Modernization Program (HPCMP) resources by verifying their credentials via SecurID-based Kerberos authentication tools. As shown in Figure 3, the GUI makes it easy for users to specify job-specific parameters, submit jobs, check the status of jobs, and analyze the results. Users can inspect the results through integrated visualization tools or download the results of the predicted protein complex by various criteria for further processes.

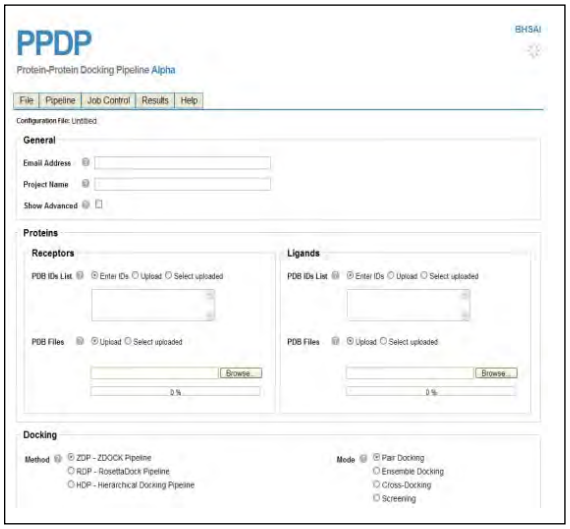

Figure 3. The Web-based graphical user interface of the PPDP

3. Results and Discussion

We applied the PPDP program to study the interaction of *Yersinia* virulence effector proteins with their cognate chaperone. The type III secretion system (T3SS) used by *Yersinia pestis* and many other Gram-negative bacteria plays an important role in the pathogenic invasion of the host cell by delivering effector proteins directly into the cytosol of the host. The delivery mechanism depends on a specific interaction of the effector protein unfolded and bound with its cognate chaperone in the bacterial cytosol (Cornelis, 2002). The interplay of disordered effector proteins and chaperones is a challenging problem to study, both experimentally and computationally. We used the PPDP to investigate the binding interaction between the disordered *Yersinia* effector protein YopE and its cognate chaperone SycE (Rodgers, 2008). Starting from compactly folded *de novo* models generated by the PSPP (Lee, 2009), a large ensemble of unfolded conformations of the chaperone binding domain of YopE ($YopE_{CBD}$) was generated with replica-exchange MD simulations (REMD). A multi-step protein docking protocol that combined ensemble docking, clustering, and structural refinement with the PPDP was used to dock the effector $YopE_{CBD}$ to its cognate SycE. The predicted $YopE_{CBD}$/SycE binding complex was in good agreement with the experimental X-ray crystal structure (Figure 4A). The results indicated that our REMD-based protein ensemble docking strategy is able to sufficiently sample the unfolded effector conformations and subsequently predict the disordered effector/chaperone binding complex. This is the first theoretical study at the atomic level of unfolded bacterial effector/chaperone interactions (Hu, 2009).

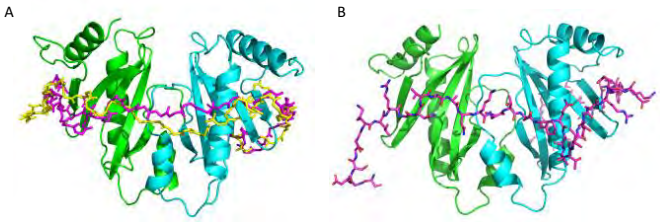

Figure 4. A: Interaction of *Yersinia pestis* virulence protein YopH (stick; yellow and magenta) with its cognate chaperone SycH (ribbon; yellow and cyan). The predicted binding mode of YopH (magenta) using PPDP agreed well with the experimental structure of the YopH/SycH complex (yellow). B: Predicted YopH/SycH complexes using PPDP. The effector YopH is shown in magenta (stick) and the chaperone SycH is shown in green and cyan (ribbon).

We further applied the REMD-based ensemble docking approach with the PPDP to predict the binding interaction of *Y. pestis* virulence effector YopH with its chaperone SycH. Despite extensive experimental studies of the protein, the crystal structure of the YopH/SycH complex has not been determined. The predicted binding complex of YopH/SycH by the PPDP, consistent with the experimental study, showed that the chaperone binding domain of effector YopH adopts a similar binding mode as YopE, which wraps around its chaperone SycH dimer in an extended, non-globular form (Figure 4B). The most significant variation was found at the N-terminal region, which exhibited large fluctuations in the predicted complex. This reiterated the challenging issue of how to efficiently deal with protein flexibility and conformational changes to improve docking performance. On the other hand, the result may also imply that the binding complex of YopH/SycH is less stable compared with YopE/SycH due to the high flexibility of the chaperone binding domain of YopH.

An interesting aspect of the bacterial effector/chaperone interaction is that, although experiments have revealed that the structure of the effector and chaperone as well as the structural binding motif involved in the interaction are very similar, the effector specifically binds to its cognate chaperone with high selectivity (Lilic, 2006). Here, we attempted to

investigate the structural basis of the binding specificity of effector/chaperone interactions using the PPDP. A challenging question is whether the scores from protein-protein docking are able to discriminate the true binding partner from a large number of non-binders. To this end, we benchmarked eight effector/chaperone binding complexes and performed protein docking using the PPDP. The structures of the chaperones and the native effector were selected from their experimental complexes in the PDB (Berman, 2002). To simplify the docking problem, only the monomer chaperone/effector interaction was considered. Therefore, two segments of the effector (namely, effector a and effector b) were docked to a chaperone separately. Figure 5 shows preliminary results of cross-docking using the ZDOCK module of PPDP. The predicted binding complexes were scored by ZRANK. Analysis of the top-scored models showed that the effectors bound to the chaperones in a similar manner with the conserved β-motif interactions. This is consistent with experimental observations, indicating that the PPDP was able to generate the correct conformations of effector/chaperone binding. However, among these eight binding complexes, only four of them were correctly predicted as true binders by ZRANK scores (YopE/SycE, SptP/SicP, SicA/InvB, and YscE/YscD). The remaining predicted complexes (YopNa/SycN, YopNbYscB, YscM2/SycH, and YopD/SycD) were scored with weak binding affinities compared with other non-binders. Further studies using the PPDP with optimized docking protocol and different post-processing approaches such as clustering and MM-GBSA scoring are ongoing.

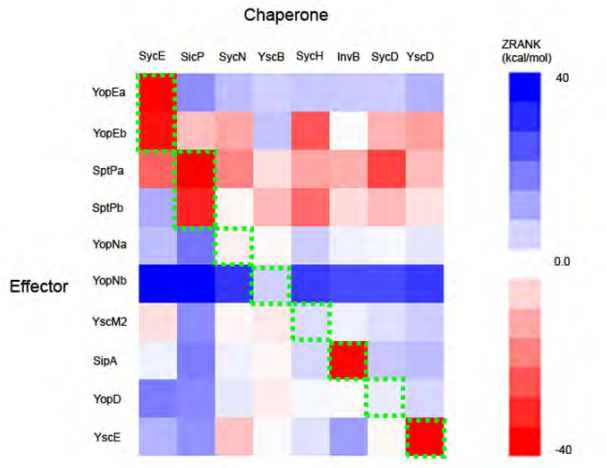

Figure 5. Prediction of bacterial virulence effector/chaperone interactions using the PPDP. The color schema according to the predicted score by ZRANK is as follows: red = lowest score (highest binding affinity) and blue=highest score (lowest binding affinity). Green boxes represent the native binding partner of the effector with its cognate chaperone (YopE/SycE, SptP/SicP, YopN/SycN, YscM2/SycH, YopD/ScyD, and YscE/YscD).

4. Conclusions

In summary, we developed an efficient protein-protein docking pipeline that integrates a variety of protein docking programs and structural refinement techniques for systematic structural prediction of protein complexes. The PPDP program has been used for studies of large-scale PPI networks of biodefense-related organisms at the BHSAI and supports various biodefense-related projects sponsored by the DoD. We applied the PPDP to study the bacterial virulence effector/chaperone interaction. As a preeminent example, the success of our PPDP approach to predict the structural "disordered-to-order" transition associated with an effector protein binding to its chaperone opens up the possibility of studying the underlying binding specificity of chaperone/effector interactions and devising possible strategies for interfering with T3SS transport.

Acknowledgments

This work was sponsored by the US Department of Defense High Performance Computing Modernization Program (HPCMP), under the High Performance Computing Software Applications Institutes (HSAI) Initiative. The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the US Army or of the US Department of Defense.

References

Andrusier, N., E. Mashiach, R. Nussinov, and H.J. Wolfson, "Principles of flexible protein-protein docking." *Proteins*, 73, pp. 271–289, 2008.

Berman, H.M., T. Battistuz, T.N. Bhat, W.F. Bluhm, and C. Zardecki, "The Protein Data Bank." *Acta Crystallogr D Biol Crystallogr*, 58, pp. 899–907, 2002.

Case, D.A., T. Darden, and P.A. Kollman, *AMBER 10*, University of California, San Francisco, 2008.

Chen, R., L. Li, and Z.P. Weng, "ZDOCK: An initial-stage protein-docking algorithm." *Proteins*, 52, pp. 80–87, 2003.

Cornelis, G.R., "Yersinia type III secretion: send in the effectors." *Journal Cell Biology*, 158, pp. 401–408, 2002.

Feig, M., J. Karanicolas, and C.L. Brooks, "MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology." *Journal Molecular Graphic Modelling*, 22, pp. 377–395, 2004.

Gray, J.J., S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C.A. Rohl, and D. Baker, "Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations." *Journal of Molecular Biology*, 331, pp. 281–299, 2003.

Lee, M.S., R. Bondugula, V. Desai, N. Zavaljevski, I.C. Yeh, A. Wallqvist, and J. Reifman, "PSPP: a protein structure prediction pipeline for computing clusters." *PLoS One* 4, e6254, 2009.

Liang, S., S. Liu, C. Zhang, and Y.Q. Zhou, "A simple reference state makes a significant improvement in near-native selections from structurally refined docking decoys." *Proteins*, 69, pp. 244–253, 2007.

Lilic, M., M. Vujanac, and C.E. Stebbins, "A common structural motif in the binding of virulence factors to bacterial secretion chaperones." *Molecular Cell*, 21, pp. 653–664, 2006.

Hu, X., M.S. Lee, and A. Wallqvist, "Interaction of the disordered Yersinia effector protein YopE with its cognate chaperone SycE." *Biochemistry*, 48, pp. 11158–11160, 2009.

Pierce, B. and Z.P. Weng, "ZRANK: Reranking protein docking predictions with an optimized energy function." *Proteins*, 67, pp. 1078–1086, 2007.

Ritchie, D.W., "Recent progress and future directions in protein-protein docking." *Current Protein & Peptide Science*, 9, pp. 1–15, 2008.

Rodgers, L., A. Gamez, R. Riek, and P. Ghosh, "The type III secretion chaperone SycE promotes a localized disorder-to-order transition in the natively unfolded effector YopE." *Journal of Biological Chemistry*, *283*, pp. 20857–20863, 2008.

Zhang, C., S. Liu, and Y. Zhou, "Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential." *Protein Science*, 13, pp. 391–399, 2004.

Vajda, S., I.A. Vakser, M.J. Sternberg, and J. Janin, "Modeling of protein interactions in genomes." *Proteins*, 47, pp. 444–446, 2002.